Hierarchical Transformers with Delta-Attention for Multimodal Fusion

Kunjal Panchal, 32126469, kpanchal@umass.edu

1 Problem Statement.

In human-computer interaction; the tasks like emotion recognition, sentiment analysis, personality trait recognition, audio source localization, audio diarization are very important for the use cases like product and advertisement market value analysis, personal assistant to the people with visible or invisible disabilities and such. The task of emotion recognition has attracted attention of the machine learning community [4, 8]. Emotions can be detected from facial expressions, from speech patterns, and from words uttered. The issue with analysis of any one mode of expression (Visual, Verbal or Vocal) is that lot of contextual information can get lost [2, 15]. E.g., a sarcastic remark like "That sounds great" can be uttered with a smirk, in an angry or indignant voice. If we only analyze the verbal modality, we would miss the context of the situation (imparted by visuals and vocals). This asks researchers to inspect multiple modalities to get a thorough understanding of the cross-modal dependencies and temporal context of the situation to analyze the emotion. [18]

The current problems regarding human multimodal behaviors during social interactions can have consequences on mental health and invisible disabilities. Along with the observation of health impacts of social anxiety, I also want to explore the task of image captioning which deals with looking at a video or an image and describing the context of the content verbally. These kinds of systems help the visually impaired people and can be a great addition to the personal assistants available in the market like Alexa, Siri, Cortana, Google Assistant and such.

2 Related Work.

To combine the most widely observable modalities: visuals, vocals and verbals; there are different ways proposed like hidden markov models [19] where Gaussian mixture models use hand-crafted features, fine Gaussian support vector machines [5] used with radial basis function kernel for non-linear classification , deep belief network [5,9] for depth level

feature extraction of physiological features. The deep learning models provided much better accuracy and understanding. Graph fusion based network [14] where the powerset of all the modalities create a graph structure and each representation of a node fuses with all others; clearly, they do not scale well with the increament in modalities, tensor fusion based network [18]; a promising path, current research is going on in lowering the amount of tensor computations, LSTMs + CNNs [6, 13, 17, 22, 25, 26] are the most popular but suffers from the lack of long-range dependencies and their unaligned nature as discussed further in the next section.

3 Novelty.

For all the approaches taken so far, the below mentioned missing gaps will be addressed in this project. In RNNs, looking at the entire sequence of data and outputting the last, single hidden state as its representation, often turns into the model forgetting information way past the current timestamp. They don't preserve the *long-range dependencies* [21] due to this bottleneck. Second, data can be unaligned across modalities. A frowning face may relate to a pessimistic word spoken in the past. It's not always between current word and current expression. And thus, third related point is, neutral expressions are quite idiosyncratic [7]. Some people may always look angry given their facial configuration. This raises the need for *delta attention* [25], we need to take cross-modal context plus the temporal context within each individual modality to negate the effect of "monotonous-across-the-time" features.

And lastly, the mechanism of fusing the modalities itself is not clearly figured out, the papers focusing on late fusion [11, 16] fall short in making any cross-correlation between the effect of each modality on the classification, generally early fusion methods [10] have been proven superior, this project focuses on building a hierarchical transformer architecture where each modality has some representation vectors inferred by delta self-attention unimodal modules, which will be fused together with a deep cross correlation analysis [DCCA] [12, 20] module and finally processed with a cross-attention module to classify multiple modalities into 6 emotion classes: happy, sad, anger, surprise, disgust, and fear.

4 Methodology.

This project will experiment with transformers like BERT [3] and auto-regressive transformers like XLNet [24], from Huggingface [23], where each feature has a vector representing its dependency across all other features in a sequence. This is ideal for the verbal modality, because a big language model trained on 33B tokens will have sufficient knowledge on textual entailment and contradiction. For verbal modality, the plan is to fine-tune a language model with a sequence classification head on top to learn the long-range dependencies across the sequence. To integrate it with the representations from the other two modalities, we shall remove the classification head from this model and use the fine-tuned hidden layer representations.

Reader must note that the dataset we are using, CMU-MOSEI (see section 5 for the details), comes with a utility to align the computational sequences of all the modalities according to the time-interval. This time interval information will be utilized to set the delta-attention window, which will be the interval itself. Because a sentence sequence can have both positive and negative, extreme and mild sentiments; the use of these intervals can help classify the sentence snippets with the local context derived from the emotions expressed in that specific timeframe.

While for visuals (RGB images) and vocals (audio mel-frequency spectrograph), a delta-attention (fixed-window) based LSTM will be a good starting point. These can also be trained individually with a softmax head and later the hidden layer representations can be extracted for cross-modality attention fine-tuning.

Once these three representations are learned, which we can call as self-attended to capture the long-range dependencies and delta-attention, they can be fused together with a DCCA module [20] to get a cross-modal representation. This representation will be finally fed into a cross-attention transformer decoder module alongside the skip connections from individual representations. Finally a fully connected layer connecting as the bridge between a softmax layer and the cross-attention module can classify the emotion classes. These hypotheses on adding a delta-attention to fuse the modalities with a correlation analysis will be proved fair if we can achieve a good F1 score on the classification task, which can compete with the research done without the use of delta-attention [18, 21].

5 Data Sets.

Multimodal Opinion Sentiment and Emotion Intensity [CMU-MOSEI][1] [1] is the largest and the latest dataset of sentence level sentiment analysis and emotion recognition. It contains more than 65 hours of annotated video from more than 1000 speakers and 250 topics. It has around 23k samples, each with around 1000 text features, 300 visual frames, and 150 acoustic features. Most of the transformer based papers have been benchmarked on this dataset, because of its size and the variety in "in-the-wild" emotions. Thus, this project will also use this dataset, to be able to compare the results fairly.

6 Future Directions.

Many potential and intriguing research topics for the field of multimodal fusion, includes:

1. use of finer granularity of modalities (subword, character etc),

[1] https://github.com/A2Zadeh/CMU-MultimodalSDK

2. use of different attention mechanisms to cover long-range dependencies over and across the modalities like building a part of speech trees to capture the relation between nouns and adjectives as in aspect based sentiment analysis,

3. use of generative adversarial networks to generate and discriminate between fake and real features (can detect sarcasm if the model can efficiently remove the fake positive verbal connotation),

4. variable window delta-attention,

5. handling larger dimension-ed data efficiently across the modalities,

6. inferring missing features of a modality with the help of present modalities,

7. inclusion of one more modality like "non speech voices" like cry, whimper; and,

8. addressing the distinction between the timescales of concepts of emotion and mood.

7 Applications.

As discussed in the Section 1, the concept of multimodal analysis/classifiers can help people with invisible disabilities who don't find themselves expressing their ideas and feelings efficiently. Context is the key. A personality trait detector which can gauge confidence or nervousness can help with a person's social anxiety. Many AI systems nowadays require a holistic view of the surroundings, faces can lie and words can be misinterpreted, but a human brain can most of the time see through the pretense, and so will the future AI, with the help of contextual multimodal models.

References

[1] Amir Ali Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[2] T. Baltrusaitis, C. Ahuja, and L. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587, 2011.

[5] Mohammad Mehedi Hassan, Md. Golam Rabiul Alam, Md. Zia Uddin, Shamsul Huda, Ahmad Almogren, and Giancarlo Fortino. Human emotion recognition using deep belief network architecture. *Information Fusion*, 51:10 – 18, 2019.

[6] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and - specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery.

[7] Verena Heusser, Niklas Freymuth, Stefan Constantin, and Alex Waibel. Bimodal speech emotion recognition using pre-trained language models, 2019.

[8] Yunxin Huang, Fei Chen, Shaohe Lv, and Xiaodong Wang. Facial expression recognition: A survey. *Symmetry*, 11(10), 2019.

[9] Yingying Jiang, Wei Li, M. Shamim Hossain, Min Chen, Abdulhameed Alelaiwi, and Muneer Al-Hammadi. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*, 53:209 – 221, 2020.

[10] Sushant Kafle, Cecilia Ovesdotter Alm, and Matt Huenerfauth. Fusion Strategy for Prosodic and Lexical Representations of Word Importance. In *Proc. Interspeech 2019*, pages 1313–1317, 2019.

[11] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[12] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using deep canonical correlation analysis. In *International Joint Conference on Neural Networks (IJCNN)*, 2019.

[13] Sijie Mai, Haifeng Hu, and Songlong Xing. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 481–492, Florence, Italy, July 2019. Association for Computational Linguistics.

[14] Sijie Mai, Haifeng Hu, and Songlong Xing. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[15] Catherine Marechal, Dariusz Mikolajewski, Krzysztof Tyburek, Piotr Prokopowicz, Lamine Bougueroua, Corinne Ancourt, and Katarzyna Wegrzyn-Wolska. *Survey on AI-Based Multimodal Methods for Emotion Detection*, pages 307–324. Springer International Publishing, Cham, 2019.

[16] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, and V. Makarenkov. A hybrid latent space data fusion method for multimodal emotion recognition. *IEEE Access*, 7:172948–172964, 2019.

[17] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. Extending long short-term memory for multi-view structured learning. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 338–353, Cham, 2016. Springer International Publishing.

[18] Saurav Sahay, Eda Okur, shachi H Kumar, and Lama Nachman. Low rank fusion based transformers for multimodal sequences. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 29–34, Seattle, USA, July 2020. Association for Computational Linguistics.

[19] Gaurav Sahu. Multimodal speech emotion recognition and ambiguity resolution. *CoRR*, abs/1904.06022, 2019.

[20] Imran Sheikh, Rupayan Chakraborty, and Sunil Kumar Kopparapu. Audio-visual fusion for sentiment classification using cross-modal autoencoder. In *Visually Grounded Interaction and Language (ViGIL) NeurIPS 2018 Workshop*, 2018.

[21] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy, 7 2019. Association for Computational Linguistics.

[22] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *CoRR*, abs/1811.09362, 2018.

[23] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[24] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019.

[25] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. *CoRR*, abs/1802.00927, 2018.

[26] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. *CoRR*, abs/1802.00923, 2018.